UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Tarea 2: Modelos de regresión

Aprendizaje Automático

Nombre: Sergio Andrés Elizondo Rodríguez

Grupo: 003

Maestro: José Anastacio Hernández Saldaña

Sábado 20 de Julio del 2024

Introducción

NSL-KDD es un conjunto de datos ampliamente utilizado en la investigación de sistemas de detección de intrusiones (*IDS*, por sus siglas en inglés). Está compuesto por registros de tráfico de red, con características que describen cada conexión, y su etiqueta correspondiente, la cual indica si la conexión es normal o se trata de un ataque.

El análisis realizado en el presente reporte busca evaluar varios modelos de regresión entrenados para predecir la variable objetivo del conjunto, esto con el fin de identificar el modelo que mejor represente los datos. Para ello, se realiza un ajuste de hiperparámetros y se comparan múltiples técnicas de regresión.

Procesamiento previo

Antes de aplicar los modelos, se realizó un procesamiento previo para garantizar que las variables sometidas a análisis fueran las adecuadas; en particular, se aplicaron dos tratamientos diferentes:

- **Variables categóricas:** Representan categorías o grupos discretos (como el tipo de conexión). Se convirtieron en representaciones binarias utilizando *OneHotEncoder* para poder usarlas en los modelos de regresión.
- **Variables numéricas:** Representan datos cuantitativos (como las métricas de tráfico de red). Se normalizaron para que fueran compatibles con todos los modelos.

Modelos

Para garantizar una evaluación robusta, se implementó una búsqueda de hiperparámetros en cuadrícula (*grid search*) con validación cruzada (*KFold*), seleccionando los parámetros que maximicen el coeficiente de determinación R^2. Se utilizó un muestreo de 400 elementos para reducir el tiempo de búsqueda.

A continuación, se presentan los modelos aplicados y los resultados obtenidos:

- **Regresión lineal.** Este modelo mostró un desempeño significativamente bajo, con un coeficiente de determinación R^2 de 0.046. Esto sugiere que el modelo lineal no logra capturar adecuadamente la variabilidad de los datos.
- **Regresión *Ridge*.** Obtuvo un R^2 de 0.3955 utilizando $\alpha = 1$. Aunque el rendimiento es mejor que el de la regresión lineal básica, sigue siendo relativamente bajo.
- **Regresión *Lasso*.** Obtuvo un R^2 de 0.6057 para $\alpha = 0.1$. Este modelo presenta una notable mejoría con respecto a otras regresiones lineales, lo cual indica que la regularización ayudó a mejorar el ajuste (sin llegar al sobreajuste) al reducir la complejidad del modelo.
- **Regresión polinómica.** Los modelos de regresión polinómica mostraron resultados extremadamente negativos. En el caso de la regresión básica y *Ridge*, se obtuvieron valores de R^2 menores a -190,000; Lasso obtuvo un mejor rendimiento ($R^2 = -0.0012$), el cual sigue siendo muy bajo en comparación con otros modelos. Estos resultados sugieren un sobreajuste severo

(el cual *Lasso* pudo minimizar hasta cierto grado) debido a la inclusión de características polinómicas de mayor grado.

- **Árbol de decisión.** Mostró un R^2 de 0.8263 con los parámetros óptimos (profundidad máxima de 7 y una configuración para la cantidad mínima de muestras por división y hoja de 5 y 2, respectivamente). Este modelo proporciona un buen ajuste y presenta la ventaja de ser fácilmente interpretable.
- ***K* vecinos más cercanos.** Presenta un R^2 de 0.8807 tomando 7 vecinos y utilizando una ponderación por distancia. Esto indica un ajuste excelente a los datos, posiblemente por su capacidad de capturar relaciones no lineales.
- **Bosque aleatorio.** Obtuvo el mejor R^2 (0.9255) con los parámetros óptimos (200 estimadores, profundidad máxima de 20 niveles, cantidad de características proporcional a la raíz cuadrada y configuración para la cantidad mínima de muestras por división y hoja de 2 y 1, respectivamente). Se muestra que el bosque aleatorio tiene una alta capacidad para modelar la complejidad en los datos, superando a los demás modelos utilizados en este caso.

Resultados

Para evaluar el desempeño de los modelos de regresión en el conjunto de datos completo, se realizó una prueba utilizando el mejor modelo identificado en la optimización: el bosque aleatorio. A continuación, se presentan las métricas obtenidas:

- **R^2:** 0.315
- **Error cuadrático promedio:** 0.168
- **Raíz del error cuadrático promedio:** 0.4098
- **Error absoluto promedio:** 0.2152

Estos resultados indican que el bosque aleatorio con los hiperparámetros óptimos muestra una capacidad moderada para explicar la variabilidad de los datos de detección de intrusiones. El valor de R^2 sugiere que aproximadamente 31.5% de la variabilidad es explicada por el modelo. Aunque el valor no es particularmente alto, indica que el modelo captura una parte significativa de la estructura de los datos.

Bibliografía

- **Base de datos:** NSL-KDD. GitHub. Recuperado el 21 de julio de 2024, de https://github.com/topics/nsl-kdd